УДК 519.233.5:61

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ С ИСПОЛЬЗОВАНИЕМ ПАКЕТА СТАТИСТИЧЕСКИХ ПРОГРАММ STATA

© 2014 г. **Т. Н. Унгуряну, *А. М. Гржибовский**

Северный государственный медицинский университет, г. Архангельск
*Норвежский институт общественного здравоохранения,
г. Осло, Норвегия

В настоящей работе рассматриваются основные принципы применения корреляционного анализа в биомедицинских исследованиях. Приводятся практические примеры анализа с расчетом коэффициентов корреляции Пирсона и Спирмена как вручную, так и с помощью пакета статистических программ STATA. Разбираются основные условия применения корреляционного анализа и правила представления результатов в научных публикациях. Материал дает общие представления о корреляционном анализе и не заменяет изучения специализированной литературы.

Ключевые слова: корреляционный анализ, коэффициент Пирсона, коэффициент Спирмена, STATA

Корреляционный анализ является одним из самых популярных методов обработки данных в российских биомедицинских публикациях, однако не всегда он применяется корректно. В связи с этим мы представляем основные принципы применения корреляционного анализа, расчет коэффициентов корреляции вручную и с помощью программного пакета STATA, а также разбираем основные условия применения данного метода.

Корреляционный анализ определяет степень связи между переменными. Направление корреляционной связи может быть прямое (положительное) или обратное (отрицательное):

• При прямой связи с увеличением значений первого признака увеличиваются значения второго признака, а с уменьшением значений первого признака происходит уменьшение значений второго.

• При обратной связи значения первого признака изменяются под воздействием второго, но в противоположном направлении по сравнению с изменением второго признака.

Сила связи (степень, теснота связи) – степень сопряженности признаков, широта варьирования значений каждого из них при изменении величины другого. Связь считается сильной, когда каждой величине одного признака соответствуют такие величины другого признака, которые сравнительно мало отклоняются от своей средней, теснее группируются вокруг нее, и наоборот, связь называется слабой (менее тесной), если значениям одного признака соответствуют значительные колебания значений второго. Сила связи не зависит от ее направленности и определяется по абсолютному значению коэффициента корреляции (табл. 1).

Таблица 1

Количественные критерии оценки силы корреляционной связи [2]

Характер связи	Величина коэффициента корреляции	
	Прямая (+)	Обратная (−)
Отсутствует	0,0	0,0
Слабая	от 0,01 до 0,29	от −0,01 до −0,29
Средняя	от 0,3 до 0,69	от −0,3 до −0,69
Сильная	от 0,7 до 0,99	от −0,7 до −0,99
Полная	1,0	−1,0

Полная (функциональная) связь – связь, при которой определенному значению одного признака соответствует одно и только одно значение другого признака. Функциональная связь проявляется во всех случаях наблюдения и для каждой конкретной единицы исследуемой

совокупности. Этот вид связи характерен для объектов, являющихся точкой приложения точных наук. В медико-биологических исследованиях функциональная связь встречается очень редко, так как объекты этих исследований имеют большую индивидуальную вариабельность.

При изучении корреляционной связи чаще всего используют численные критерии или коэффициенты.

Линейный коэффициент корреляции Пирсона (r_p) используется для измерения тесноты связи между двумя количественными признаками X и Y. Расчет коэффициента может производиться только при соблюдении следующих условий:

- Обе переменные являются количественными и непрерывными.
- Как минимум один из признаков (а лучше оба) имеет нормальное распределение.
- Зависимость между переменными носит линейный характер.
- Гомоскедастичность (вариабельность одной переменной не зависит от значений другой переменной).
- Независимость участников исследования друг от друга.
- Парность наблюдений (признак X и признак Y изучаются у одних и тех же участников исследования).

Расчет коэффициента корреляции Пирсона. Для расчета коэффициента значения переменных X и Y располагают в ряд, в котором каждой величине X соответствует определенная величина Y. Затем рассчитывают средние арифметические значения для каждой переменной $\overline{X}$ и $\overline{Y}$ соответственно. Далее следует найти отклонения каждого значения X и Y от соответствующей средней величины и перемножить отклонения для X и Y между собой. Таким образом, получаем числитель для формулы расчета коэффициента Пирсона. Для знаменателя необходимо рассчитать стандартные отклонения для X и Y (s_x и s_y). Полученные промежуточные величины подставляются в формулу расчета коэффициента (r_p):

$$r_p = \frac{\sum(X - \overline{X}) \cdot (Y - \overline{Y})}{(n - 1) \cdot s_X \cdot s_Y}$$

где: X – значения независимой переменной, Y – значения зависимой переменной, $\overline{X}$ – среднее арифметическое значение переменной X, $\overline{Y}$ – среднее арифметическое значение переменной Y, s_x и s_y – стандартные отклонения для переменных X и Y, n – количество пар наблюдений.

Для оценки статистической значимости выявленной взаимосвязи между переменными необходимо провести сравнение расчетного значения коэффициента Пирсона с критическим значением, взятым из таблицы. Если расчетное значение r_p равно или превышает критическое значение $r_{p0,05}$, то H_0 отвергается и делается вывод о том, что коэффициент корреляции статистически значимо отличается от нуля ($p < 0,05$).

Пример. Во время мониторинга проводилось измерение органолептических и санитарно-химических показателей водопроводной воды на 12 водоколонках города. Для оценки влияния на цветность воды исследовано содержание железа (табл. 2). С помощью корреляционного анализа необходимо выявить наличие зависимости между цветностью (Y) и концентрацией железа (X) в водопроводной воде.

Таблица 2

Расчет коэффициента корреляции Пирсона

№	X	Y	$X - \overline{X}$	$Y - \overline{Y}$	$(X - \overline{X}) \times (Y - \overline{Y})$	$(X - \overline{X})^2$	$(Y - \overline{Y})^2$
1	0,08	15	−0,17	−7,5	1,275	0,0289	56,25
2	0,15	15	−0,1	−7,5	0,75	0,01	56,25
3	0,19	20	−0,06	−2,5	0,15	0,0036	6,25
4	0,29	21	0,04	−1,5	-0,06	0,0016	2,25
5	0,23	21	−0,02	−1,5	0,03	0,0004	2,25
6	0,25	22	0	−0,5	0	0	0,25
7	0,27	23	0,02	0,5	0,01	0,0004	0,25
8	0,23	24	−0,02	1,5	−0,03	0,0004	2,25
9	0,24	25	−0,01	2,5	−0,025	0,0001	6,25
10	0,31	25	0,06	2,5	0,15	0,0036	6,25
11	0,29	26	0,04	3,5	0,14	0,0016	12,25
12	0,41	33	0,16	10,5	1,68	0,0256	110,25
				Сумма	4,07	0,0762	261

Расчеты показали, что средние арифметические значения цветности и концентрации железа по всем 12 водоколонкам составили: $\overline{X} = 0,25$ градусов и $\overline{Y} = 22,5$ мг/л соответственно, а стандартное отклонение: $s_X = 0,083$ градуса и $s_Y = 4,87$ мг/л.

$$r_p = \frac{4,07}{(12 - 1) \cdot 0,083 \cdot 4,87} = \frac{4,07}{4,44} = 0,91$$

Расчетное значение коэффициента корреляции Пирсона (r_p) в данном примере оказалось равно 0,91. Для оценки нулевой гипотезы необходимо расчетное значение критерия (r_p) сравнить с табличным значением критерия. Из таблицы критических значений критерия корреляции Пирсона для n = 12 и уровня статистической значимости 0,001 критическое значение r_p составляет 0,823. Так как расчетное значение больше критического, выявленная взаимосвязь между содержанием железа и цветностью водопроводной воды является статистически значимой. Кроме того, по величине коэффициента корреляции и знаку, с которым он получился, можно судить о силе и направлении связи. В данном примере коэффициент корреляции равен +0,91, что свидетельствует о прямой и сильной зависимости, то есть чем выше содержание железа, тем выше цветность воды.

Расчет коэффициента корреляции Пирсона в STATA. Сначала необходимо проверить условия применения коэффициента. Для проверки нормальности распределения переменных следует в меню Statistics выбрать Summaries, tables, and tests → Distributional plots and tests → Shapiro-Wilk normality test. В поле Variables можно перенести сразу обе переменные

Colour (цветность) и Iron (железо) (рис. 1). Результаты теста Shapiro-Wilk показали, что обе переменные имеют нормальное распределение (рис. 2).

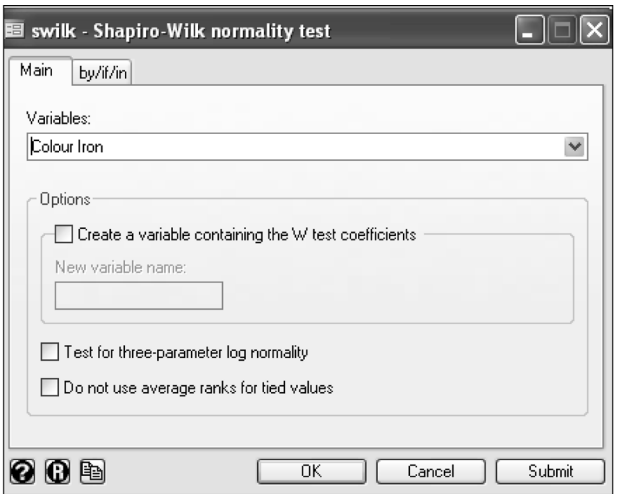

Рис. 1. Диалоговое окно для расчета теста Shapiro-Wilk

Shapiro-Wilk W test for normal data

Variable |	Obs	W	V	z	Prob>z
Colour	12	0.94222	0.965	-0.069	0.52731
Iron	12	0.97203	0.467	-1.483	0.93090

Рис. 2. Результаты теста Shapiro-Wilk

Для определения линейности связи между переменными следует построить скатерограмму. Для этого в меню Graphics нужно выбрать Twoway graph (scatter, line, etc). Появится диалоговое окно twoway – Twoway graphs, в котором нужно нажать на Create → выбрать Basic plots → Scatter. В поле Y variable следует выбрать зависимую переменную Colour, а в поле X variable – независимую Iron (рис. 3). На рис. 4 видно, что зависимость между переменными X и Y носит линейный характер.

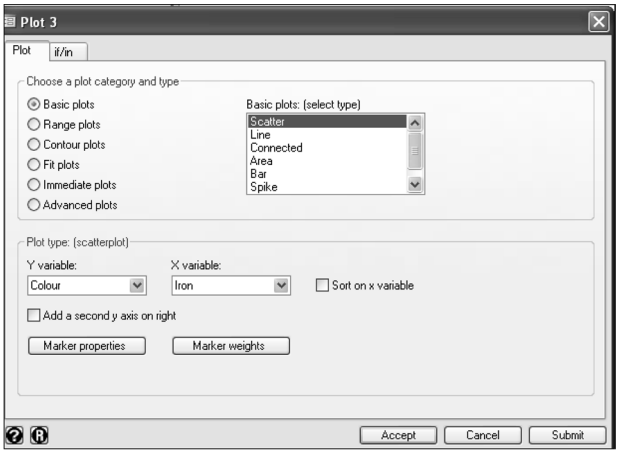

Рис. 3. Диалоговое окно для построения скатерограммы (Scatter)

Для расчета коэффициента корреляции Пирсона необходимо в меню Statistics выбрать Summaries, tables, and tests → Summary and descriptive statistics →

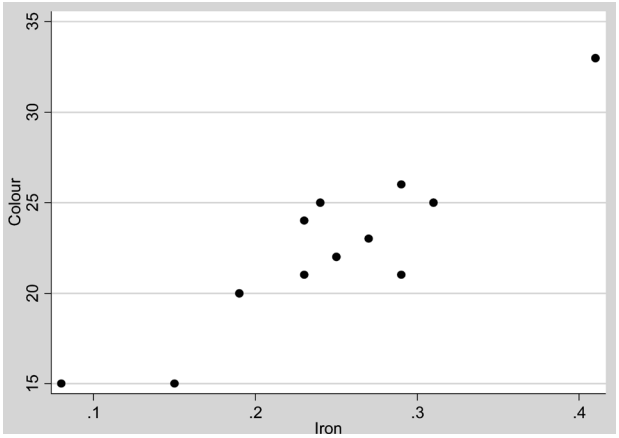

Рис. 4. Скатерограмма

Pairwise correlations. Откроется диалоговое окно pwcorr – Pairwise correlations of variables, в котором в поле Variable следует перенести обе переменные Colour и Iron и обязательно нужно поставить галочку рядом с Print significance level for each entry, для того чтобы в результатах отобразился уровень статистической значимости.

Результаты расчета коэффициента корреляции Пирсона показали, что между железом и цветностью существует сильная прямая взаимосвязь ($r_p = 0{,}9144$). Согласно данным самой нижней строки (0,0000) коэффициент корреляции статистически значимо отличается от нуля ($p < 0{,}001$).

```
. pwcorr Colour Iron, sig

             |  Colour    Iron
-------------+------------------
      Colour |  1.0000
        Iron |  0.9144   1.0000
             |  0.0000
```

Рис. 5. Результаты расчета коэффициента корреляции Пирсона в STATA

Коэффициент корреляции рангов Спирмена (r_s). Непараметрический коэффициент корреляции рангов Спирмена используется, когда распределение значений количественных переменных не соответствует нормальному распределению или если необходимо оценить связь между качественными (порядковыми) и количественными признаками или только между порядковыми признаками.

Расчет коэффициента Спирмена. Сначала нужно составить два ряда из парных сопоставляемых признаков, обозначив первый и второй ряд соответственно X и Y. При этом представить первый ряд признака в убывающем или возрастающем порядке, а числовые значения второго ряда расположить напротив того значения первого ряда, которым они соответствуют. Затем значения первой и второй переменных заменяют порядковым номером (рангом). При этом числовым значениям второго признака ранги должны присваиваться в том же порядке, какой был принят

при раздаче их величинам первого признака. При одинаковых величинах признака в ряду ранги следует определять как среднее число из суммы порядковых номеров этих величин. После ранжирования определяют разности рангов (d) между ранговыми номерами X и Y, возводят их в квадрат (d^2) и суммируют. Полученную сумму квадратов разности рангов (Sd^2) подставляют в формулу расчета коэффициента корреляции Спирмена (r_s):

$$r_s = 1 - \frac{6 \cdot \sum d^2}{n^3 - n}$$

где n – число сравниваемых пар.

Для оценки статистической значимости выявленной взаимосвязи между переменными необходимо провести сравнение расчетного значения коэффициента Спирмена с критическим значением, взятым из таблицы. Если расчетное значение r_s равно или превышает критическое значение $r_{s0,05}$, то H_0 отвергается и делается вывод о том, что коэффициент корреляции статистически значимо отличается от нуля ($p < 0,05$).

Пример. Во время мониторинга проводилось измерение санитарно-химических показателей водопроводной воды на 12 водоколонках города. Для оценки влияния на жесткость воды исследовано содержание кальция (табл. 3). С помощью корреляционного анализа необходимо выявить наличие зависимости между жесткостью (Y) и концентрацией кальция (X) в водопроводной воде.

Таблица 3

Расчет коэффициента корреляции Спирмена

№	X	Y	Ранг X	Ранг Y	d	d^2
1	0,36	6,4	1,5	8	−6,5	42,25
2	0,36	4	1,5	2,5	−1	1
3	0,38	3,2	3	1	2	4
4	0,4	4,4	4	4	0	0
5	0,48	6,8	5	9	-4	16
6	0,5	5	7	6	1	1
7	0,5	4,8	7	5	2	4
8	0,5	4	7	2,4	4,6	21,16
9	0,6	6	9	7	2	4
10	0,7	7,4	10	10	0	0
11	0,9	9,2	11	11	0	0
12	1,15	12	12	12	0	0
					Сумма	93,41

$$r_s = 1 - \frac{6 \times 93,41}{12^3 - 12} = 1 - \frac{560,46}{1716} = 1 - 0,33 = 0,67.$$

Расчетное значение коэффициента корреляции Спирмена (r_s) в данном примере оказалось равно 0,67. Для оценки нулевой гипотезы необходимо расчетное значение критерия (r_s) сравнить с табличным значением критерия. Из таблицы критических значений критерия корреляции Спирмена для n = 12 и уровня статистической значимости 0,05 критическое значение r_s составляет 0,58. Так как расчетное значение больше критического, выявленная взаимосвязь между содержанием кальция и жесткостью водопроводной воды является статистически значимой. Кроме того, по величине коэффициента корреляции и знаку, с которым он получился, можно судить о силе и направлении связи. В данном примере коэффициент корреляции равен +0,67, что свидетельствует о прямой и средней зависимости, то есть чем выше содержание кальция, тем выше жесткость воды.

Расчет коэффициента корреляции Спирмена в STATA. Проверка нормальности распределения переменных с помощью теста Shapiro-Wilk показала, что переменная Кальций (Calcium) имеет нормальное распределение, а переменная Жесткость (Hardness) не подчиняется закону нормального распределения (рис. 6), поэтому для выявления взаимосвязи между двумя переменными следует использовать коэффициент корреляции Спирмена.

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
Hardness	12	0.82004	3.007	2.145	0.01597
Calcium	12	0.89706	1.720	1.057	0.14534

Рис. 6. Результаты проверки типа распределения переменных Кальций и Жесткость

Для расчета коэффициента корреляции Спирмена необходимо в меню Statistics выбрать Summaries, tables, and tests → Nonparametric tests of hypotheses → Spearman's rank correlation. Откроется диалоговое окно spearman – Spearman's rank correlation coefficients (рис. 7), в котором в поле Variable следует перенести обе переменные Calcium и Hardness и

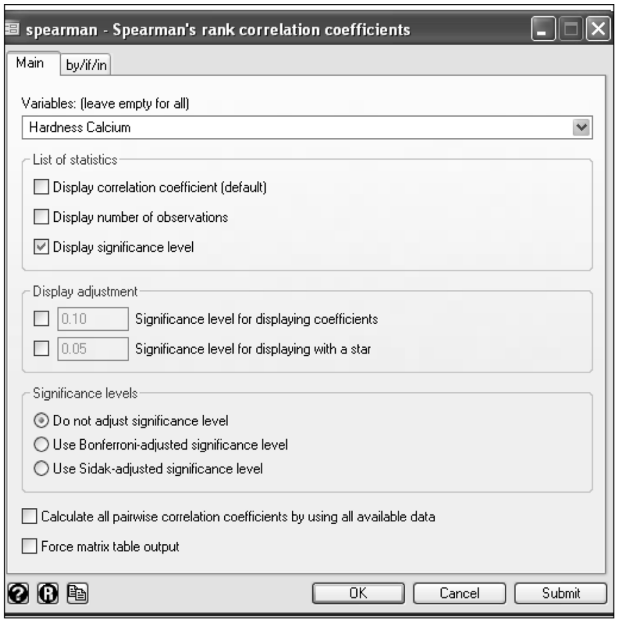

Рис. 7. Диалоговое окно для расчета коэффициента корреляции Спирмена

обязательно нужно поставить галочку рядом с Display significance level, для того чтобы в результатах отобразился уровень статистической значимости.

Результаты расчета коэффициента корреляции Спирмена показали (рис. 8), что между кальцием и жесткостью существует средней силы прямая взаимосвязь ($r_p = 0{,}6732$). Согласно данным самой нижней строки, коэффициент корреляции статистически значимо отличается от нуля ($p = 0{,}0164$).

```
. spearman Hardness Calcium, stats(p)

 Number of obs =      12
Spearman's rho =      0.6732

Test of Ho: Hardness and Calcium are independent
    Prob > |t| =      0.0164
```

Рис. 8. Результаты расчета коэффициента корреляции Спирмена в STATA

В публикациях целесообразно представлять значение коэффициентов корреляции (достаточно двух знаков после запятой), размер выборки и достигнутый уровень значимости (достаточно трех знаков после запятой), причем данная рекомендация справедлива как для коэффициента корреляции Пирсона, так и для коэффициента Спирмена. Более детальные рекомендации представления результатов корреляционного анализа представлены в [3]. В настоящее время многие зарубежные журналы рекомендуют вместо уровня значимости представлять доверительные интервалы для коэффициентов корреляции. К сожалению, программа STATA доверительные интервалы для коэффициентов корреляции не рассчитывает, но их можно рассчитать вручную с помощью формул, рассматриваемых нами в одном из предыдущих выпусков Практикума [1]. Более детальная информация об использовании пакета прикладных статистических программ STATA для проведения корреляционного анализа представлена в [4].

Список литературы

1. *Гржибовский А. М.* Корреляционный анализ // Экология человека. 2008. № 9. С. 50–60.

2. *Марченко Б. И.* Здоровье на популяционном уровне: статистические методы исследования (руководство для врачей). Таганрог : Сфинкс, 1997. 432 с.

3. *Унгуряну Т. Н., Гржибовский А. М.* Краткие рекомендации по описанию, статистическому анализу и представлению данных в научных публикациях // Экология человека. 2011. № 5. С. 55–60.

4. *Lawrence C.* Hamilton Statistics with STATA: Updated for Version 10. / Lawrence C. Hamilton. Brooks/Cole, Cengage Learning, 2009. 491 p.

References:

1. Grjibovski A. M. Correlation analysis. *Ekologiya cheloveka* [Human Ecology]. 2008, 9, pp. 50-60. [in Russian]

2. Marchenko B. I. *Zdorovje na populyatsionnom urovne: statisticheskie metody issledovaniya* [Health on a population level: statistical research methods]. Taganrog, Sfinx Publ., 1997, 432 p.

3. Unguryanu T. N., Grjibovski A. M. Brief recommendations on description, analysis and presentation of data in scientific papers. *Ekologiya cheloveka* [Human Ecology] 2011, 5, pp. 55-60. [in Russian]

4. Lawrence C. Hamilton *Statistics with STATA: Updated for Version 10*. Brooks/Cole, Cengage Learning, 2009, 491 p.

CORRELATION ANALYSIS USING STATA

T. N. Unguryanu, *A. M. Grjibovski

International School of Public Health, Northern State Medical University Arkhangelsk, Russia
**Department of International Public Health, Norwegian Institute of Public Health, Oslo, Norway*

In this paper we present general principles of correlation analysis and its use in biomedical research. Practical examples of correlation analysis are given. Calculations of Pearson's and Spearmen's correlation coefficients are presented using formulas and STATA software. Main assumptions for the use of correlation analysis are discussed as well as general principles of presentation of the results in biomedical publications. The article presents only general information about correlation analysis and does not substitute special statistical literature.

Keywords: correlation analysis, Pearson's correlation, Spearmen's correlation, STATA

Контактная информация:

Гржибовский Андрей Мечиславович – доктор медицины, профессор, старший советник Норвежского института общественного здравоохранения, г. Осло, Норвегия; Директор Архангельской международной школы общественного здоровья ГБУО ВПО «Северный государственный медицинский университет», г. Архангельск.

Адрес: Nasjonalt folkehelseinstitutt, Pb 4404 Nydalen, 0403 Oslo, Norway

Тел.: +47 22048319, +47 45268913

E-mail: andrej.grjibovski@gmail.com